Towards Semantic Annotation for Scientific Datasets

Alsayed Algergawy, Hamdi Hamed, Sven Thiel, and Birgitta König-Ries

Heinz-Nixdorf Chair for Distributed Information Systems,
Institute for Computer Science, University of Jena, Germany
{alsayed.algergawy|hamdi.hamed|sven.thiel|birgitta.koenig-ries@uni-jena.de}

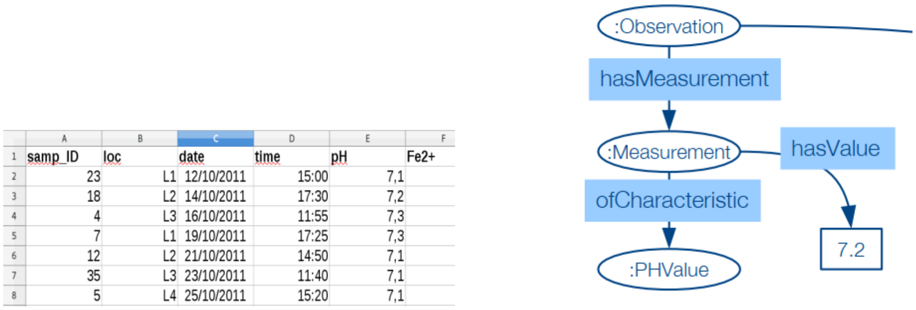
Abstract Semantic Web resources provide essential demands to support dataset search, findability and interoperability. This need becomes a necessity in long-running and large scale collaborative projects, as it affords semantic languages to enrich interpretation of commonly used terms within the project. To end this, in this paper, we introduce a new tool that supports the (semi-)automatic annotation of scientific datasets collected within the framework of CRC AquaDiva[1]. To validate the powerful data annotation, we demonstrate the deployment of data annotation to enhance dataset search and analysis.

1 Introduction

As the complexity and amount of data collected within the context of long-running and large scale projects, e.g. the Collaborative Research Center (CRC) AquaDiva, there is a growing need to make use of semantic web to standardize data. Where the annotation of scientific datasets with ontology entities provides unique opportunities for data findability and interoperability[4]. To demonstrate the importance of semantic annotation for scientific data, we introduce a simple example as shown in Fig. 1a. It represents a set of samples (*samp_ID*) collected at different locations (*loc*) and at different dates (*date*) and times (*time*). From each sample, a number of observations have been measured, such as *pH* and *Fe2+*. After storing the dataset into the associated data portal, a new user requests datasets about "observations referring to alkaline milieu", but as expected she will not get any answer. However, if the dataset is annotated by a semantic resource, e.g. an ontology as shown in Fig. 1b, the system will be able to answer the user query using a simple reasoner.

Indeed, getting the right dataset(s) as a response to the user query, a number of preprocessing tasks, including dataset annotation are required. To this end, we introduce a (semi-)automated tool that allows dataset owners to annotate their datasets during the upload process. Furthermore, the tool supports editing and modifying the existing annotations to correct and enhance the dataset

[1] http://www.aquadiva.uni-jena.de/



	A	B	C	D	E	F
1	samp_ID	loc	date	time	pH	Fe2+
2	23	L1	12/10/2011	15:00	7,1	
3	18	L2	14/10/2011	17:30	7,2	
4	4	L3	16/10/2011	11:55	7,3	
5	7	L1	19/10/2011	17:25	7,3	
6	12	L2	21/10/2011	14:50	7,1	
7	35	L3	23/10/2011	11:40	7,1	
8	5	L4	25/10/2011	15:20	7,1	

(a) Sample of a dataset

(b) Sample of semantic resource

Figure 1: Sample example

interpretation. We develop our own domain-specific ontology, called AquaDiva Ontology *ADON*, as the semantic resource used for annotation. To demonstrate the deployment of the annotation tool, we describe its use to enhance dataset search and semantic analysis.

2 Methodology

A dataset is defined as a tuple of primary data and metadata organized for a specific purpose. The primary data represents the actual data organized according to a specific structure, called *data structure*. Each data structure consists of a set of data attributes, each data attribute has a name, datatype, (optional) unit, and description. Each tuple in the primary data is a collection of data cells containing the actual data values (called *data points*) [2]. The metadata contains information about the data owner, data curators, the methodology used to produce primary data, etc. In the implementation, we are going to annotate data attributes of available datasets with corresponding concepts from the domain-specific ontology (*ADOn*). In the following, we are going to elaborate more on two main components: *ontology development* and *annotation scheme*.

2.1 Ontology development

To develop the domain specific ontology for the CRC AquaDiva, we make use of the fusion/merge strategy, where the new ontology is developed by assembling and reusing one or more (parts of) existing ontologies. To this end, we make use of the available resources in the project, such as project proposals, as well as we collected numerous research (competency) questions from the project scientists. We analyzed these resources and extracted main terms that cover the project domains, such as ecology, biology and aquatic. These terms have been used as input to the JOYCE tool[3] implemented within the project. The current

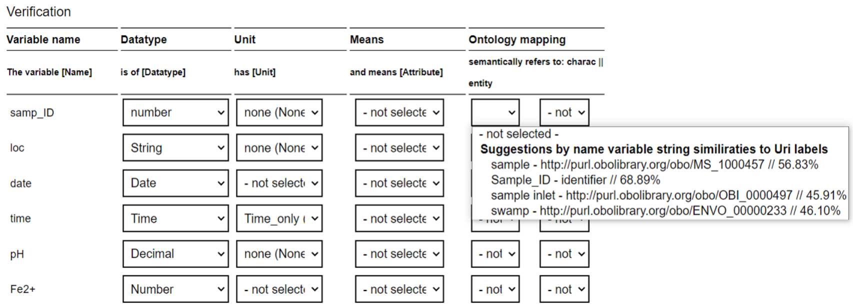
AquaDiva ontology (*ADOn*) has 78.840 axioms, 8.892 concepts, and 245 object properties.

2.2 Annotation scheme

Data annotation is the process of associating a dataset component with a concept from the *ADOn* ontology. It is possible by using this annotation to search and explore data repositories. As it has the effect to link collection of datasets in a data repository to well-defined concepts. However, annotating scientific datasets is a hard process, as it is mandatory to not only link the data attribute with a concept, but also it is required to identify the data attribute context. Consider, two datasets $\mathcal{DS}_1$ and $\mathcal{DS}_2$ stored in a data repository, e.g., the AquaDiva data repository. The first dataset $\mathcal{DS}_1$ contains `weather and soil monitoring` data. It has a data structure with 50 data attributes including *"soil temperature"* annotated with the concepts soil (`http://purl.obolibrary.org/obo/ENVO_00001998`) and characteristic temperature (`http://purl.obolibrary.org/obo/PATO_0000146`). The second dataset $\mathcal{DS}_2$ provides information about `soil moisture` in the Hainich forest. $\mathcal{DS}_2$ has a data structure with 13 data attributes. The *"mean_theta_forestbottom"* data attribute is also annotated with the concept soil and the characteristic soil moisture (`http://www.aquadiva.uni-jena.de/ad-ontology/ad-ontology.0.0/ad-ontology-characteristics.owl#SoilMoisture`). Analyzing the dataset annotation, a possible relationship of the two datasets $\mathcal{DS}_1$ and $\mathcal{DS}_2$ can be determined.

Once we have the AquaDiva ontology, we can use it to annotate datasets on the data portal. We provide two ways for the annotation:

- During dataset upload: The first option is to support the annotation of dataset during the uploading process. To this end, we collect needed information, such as data attribute' name, data types, and description (if available) to create a recommended list of concepts from the AquaDiva ontology. The list is created by computing the relatedness of a data attribute with ontology concepts by measuring the similarity between collected information and context information of ontology concepts, such as concept' label and definition. To compute this similarity, we make use of three different similarity measures, Levenstein, Jaro, and Jaccard, as each can compute the similarity from one aspect. As shown in Fig. 2, the recommended list of concepts are shown to the user (assuming that the person who uploads the dataset has enough knowledge to achieve the annotation task). The figure shows that the sample dataset shown in Fig 1 is uploading through the data portal. During the uploading process we annotate the dataset as shown in Fig. 2. For example, the data attribute "samp_ID" has a list of recommended annotation, each associated with a score. After that, the user can either select a concept from the list or select another concept from the ontology.
- After uploading dataset: In some cases, the user who uploads the dataset does not have enough knowledge to annotate the dataset, so we need to manually annotate the dataset and then we update the annotation tables.



Verification

Variable name	Datatype	Unit	Means	Ontology mapping	
The variable [Name]	is of [Datatype]	has [Unit]	and means [Attribute]	semantically refers to: charac || entity	
samp_ID	number	none (None	- not selecte		- not
loc	String	none (None	- not selecte		
date	Date	- not selecte	- not selecte		
time	Time	Time_only (	- not selecte	- not	- not
pH	Decimal	none (None	- not selecte	- not	- not
Fe2+	Number	- not selecte	- not selecte	- not	- not

- not selected -
Suggestions by name variable string similiraties to Uri labels
sample - http://purl.obolibrary.org/obo/MS_1000457 // 56.83%
Sample_ID - identifier // 68.89%
sample inlet - http://purl.obolibrary.org/obo/OBI_0000497 // 45.91%
swamp - http://purl.obolibrary.org/obo/ENVO_00000233 // 46.10%

Figure 2: Annotation process

Furthermore, the annotation tool supports editing the existing annotation

2.3 Semantic annotation deployment

To demonstrate the effectiveness and usability of semantic annotation to datasets collected in the AquaDiva project, we briefly outline two main applications:

- Semantic search: using the annotation we can move from keyword search to semantic search looking for not only the requested term but also for its content and the intended meaning for the user request. Furthermore, it supports the possibility to search through the term hierarchy.
- Dataset linking and exploration[1]: In this application, we classify a given dataset into a domain topic. With this topic, we then extract hidden links between different datasets in the repository making use of machine learning and semantic annotation.

Acknowledgments

This work has been funded by the *Deutsche Forschungsgemeinschaft (DFG)* as part of CRC 1076 AquaDiva(Projectnumber 218627073).

References

1. A. Algergawy, H. Hamed, and B. König-Ries. Towards scientific data synthesis using deep learning and semantic web. In *The Semantic Web: ESWC 2021 Satellite Events - Virtual Event, June 6-10, 2021, Revised Selected Papers*, pages 54–59, 2021.
2. J. Chamanara, M. Owonibi, A. Algergawy, and R. Gerlach. An extensible conceptual model for tabular scientific datasets. In *The Fifth International Conference on Advances in Information Mining and Management*, 2015.
3. E. Faessler, F. Klan, A. Algergawy, B. König-Ries, and U. Hahn. Selecting and tailoring ontologies with joyce, 2017.
4. N. W. Paton, J. Chen, and Z. Wu. Dataset discovery and exploration: A survey. *ACM Computing Surveys*, 56(4):1–37, 2023.